基于 CRF 和半监督学习的中文时间信息抽取

闫紫飞，姬东鸿

（武汉大学 计算机学院，湖北 武汉 430072）

摘 要：为提高文本中时间信息识别和抽取的效率，提出一种基于 CRF（条件随机场）的方法。根据时间信息表现出的一般特点，采用机器学习的方法，通过分析文本中相关词性、短语结构和上下文信息等，提取时间信息的外部特征，采用一种自训练的半监督方法，使用 CRF 进行识别和抽取。实验结果表明，该方法有效提升了时间识别的性能，在显性时间、隐性时间和总体时间上分别取得了 96.25%、88.65%和 93.97%的 *F*1 值。

关键词：条件随机场；时间抽取；时间识别；半监督；自训练

中图法分类号：TP391.1 **文献标识号**：A **文章编号**：1000-7024（2015）06-1642-05

doi：10.16208/j.issn1000-7024.2015.06.044

Exploration of Chinese temporal information extraction based on CRF and semi-supervised learning

YAN Zi-fei, JI Dong-hong

(Computer School, Wuhan University, Wuhan 430072, China)

Abstract: To improve the efficiency of extracting temporal information from the text, one method based on conditional random fields (CRF) was proposed. According to the general characteristics shown by the temporal information, the method of machine learning was adopted, by analyzing a set of linguistic features of time phrases in text such as lexical features, syntactic features and context information, while using the semi-supervised method of self-training, temporal information was recognized and extracted using CRF. Experimental results show a good performance reaching scores of 96.25%, 88.65% and 93.97% for F-measure to dominant time, recessive time and full time.

Key words: conditional random fields; temporal extraction; temporal recognition; semi-supervised; self-training

0 引 言

在自然语言中，时间是重要的组成部分，是完整理解文章语义不可或缺的要素，在信息抽取中是一个比较重要的领域。对此进行研究，可以提高信息抽取的自动化水平，对机器翻译等人工智能领域的发展有很大促进作用。SemEval 的评测中就包含时间信息的识别问题，SemEval-2013 Task 1：TEMPEVAL-3 的任务中也有时间表达及时间关系的估计。目前时间信息抽取[1]的方法主要为基于规则的方法和基于机器学习的方法，一般认为，基本的时间短语都有着清晰的结构和明显的特征，通过构建完备的规则也可以覆盖到相当部分的时间信息，因此用基于规则的方法也能够表现出比较良好的效果。然而，一般的规则在处理复杂的时间信息时，不同规则之间会有一定的冲突。此外，基于规则的方法[2]在跨语言的时候，需要做一些额外的工作，比较费时费力。近年来随着标注语料库和标注工具的完善，基于机器学习的方法由于自动化程度较高、人工干预较少、移植能力比较强，开始流行起来。时间信息有显性和隐性两种，用规则的方法对时间信息进行抽取时，隐性时间的识别效果比较差，而且此方法的可移植性也不好，针对此问题，本文用统计的方法，采用条件随机场模型，利用半监督的训练，对时间信息进行识别研究，实验结果较好。

1 研究现状

时间表达的识别是开展相关时间关系推理、时间关联

收稿日期：2014-06-26；**修订日期**：2014-08-30

基金项目：国家自然科学基金重点项目（61133012）；国家自然科学基金项目（61173062）

作者简介：闫紫飞（1985-），男，河南邓州人，硕士研究生，研究方向为自然语言处理、机器学习；姬东鸿（1967-），男，湖北武汉人，教授，博士生导师，研究方向为自然语言处理、智能信息处理、机器学习、认知语言学。E-mail：251386416@qq.com

信息获取等应用的第一步，所以它是一种基础性的工作。Zacks 认为可以通过时间的序列结构来理解事件，由此也凸显了时间识别的重要性。

在时间抽取的研究历史中，它经常作为命名实体抽取中的一部分来进行研究。在 1998 年举行的 MUC 会议上，首次将时间评测的要求加入到了命名实体识别的任务中[3]，开了时间信息抽取研究的先河。2004 年 ACE 在其子项目 TERN（time expression recognition and normalization）中详细定义了时间表达式的识别评测，不仅要求识别出时间短语，而且还要对其语义进行处理，目标是以 TIMEX2 标注作为规范[4]，分别对英文和汉语文本中的时间表达式进行识别，并进行解释。从其评测任务[5]可以看出，时间信息的抽取仍然是一个重要的研究课题。文献［6］研究了命名实体的自动识别问题，分析了规则方法和基于统计模型识别方法的优缺点，并定义了一个中文时间框架，并制定了一个规则集，开发了一个分析器 CTEMP，用于抽取和归一化中文时间短语。文献［7］分析语料的时间关系识别时用及的各语言特征，提出了基于最大熵的方法。文献［8］在时间识别问题中引入事件时间，通过复杂的语法分析和命名实体方法挖掘时间与事件的关系，但对包含多个事件的时间序列是不适用的。文献［9］在基于 CRF（条件随机场）的命名实体识别的实验中，对中文文本进行了原子切分，选取上下文特征、词性特征、词表外部特征等作为特征集来进行实体识别，取得了不错的结果。文献［10］加入语义角色特征构建特征向量，然后采用 CRF 进行识别。但是识别的效果不是太好。文献［11］比较了现在流行的各种方法，证实了 CRF 在命名实体识别领域中的良好效果。

2 分析模型

2.1 时间信息的分类

具体来讲，时间信息可分为显性时间信息和隐性时间信息。显性时间信息就是那种人们一看到就有比较明确时间概念的信息，是由人类通过自然界周期变化总结出来的一系列时间概念，如世纪、年、月、日、分、秒等。最简单的显性信息就是这些概念加上一些量词组成的，如："2007 年、4 个月、36 小时"等；然后这些简单时间信息通过任意组合，并加上一些："前"、"后"、"从…到…"等介词或方位词就构成了复杂的时间信息，如："2010 年 3 月 4 日到 7 日"、"去年冬季前后"等。另外还有一些时间专有名词也属于显性这一类，比如："春运期间"、"圣诞节"等等。可以看到，显性信息中，时间概念的特征比较明显，专有时间名词的数量有限，所以通过构建规则的方法和基于机器学习的方法都可以取得比较良好的效果。而隐性时间是诸如"树木发芽"、"直到他做完功课"等隐藏在语义之中的信息，用规则很难全面覆盖，只能利用词性、词之间的关联和上下文等信息并通过统计学习的方法来识别。

2.2 工作准备

作者在以前参与的一个项目中，基于 ontology 的工具，曾建立过完备的时间词典库，并构建了一种基于迭代的规则方法，利用词典库，对时间短语进行抽取。在对结果进行的分析中，发现此方法对显性的时间短语识别效果还不错，即使是对开放语料的测试中，召回率和准确率都可达 90％以上；而对隐性的时间的识别效果则比较差，召回率连 80％都达不到。可以看到，时间信息特别是隐性时间信息的识别性能，还是有一定的提升空间的。所以，本文中利用统计学习的方法，基于 CRF 理论，通过对文本进行分词和标注，进行特征的提取，并结合半监督的训练方法，分别对显性时间、隐性时间和总体时间进行抽取，将结果与基于迭代规则的方法进行比较。

3 基于统计的时间识别及半监督学习

3.1 CRF 介绍

如图 1 所示：以 $\boldsymbol{X}=\{x_1, x_2, \cdots\cdots, x_n\}$ 表示观测值序列，$\boldsymbol{Y}=\{y_1, y_2, \cdots\cdots, y_n\}$ 以表示隐含的状态序列，则 x_i 取决于产生它们的状态 y_{i-1}，y_i，y_{i+1}，图中的 y_1，y_2，……等状态的序列还是一个马尔科夫链。在这个图中，顶点代表一个个随机变量，顶点之间的弧代表它们相互的依赖关系，通常采用一种概率分布，比如 $p(x_1, y_1)$ 来描述，且每个状态的转移概率只取决于相邻的状态。整个条件随机场就是在给定观察序列条件下，计算整个标注序列的联合概率分布。在给定 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 序列的条件下，线性链的 CRF 定义 $\boldsymbol{Y}$ 的条件概率为

$$p(\boldsymbol{Y}\mid\boldsymbol{X},\lambda)=\frac{1}{z(\boldsymbol{X})}\exp\left(\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1},y_i,\boldsymbol{X},i)\right) \tag{1}$$

其中

$$z(\boldsymbol{X})=\sum_{j}\exp\left(\sum_{i=1}^{n}\sum_{j}f_j(y_{i-1},y_i,\boldsymbol{X},i)\right) \tag{2}$$

式（2）是归一化因子，n 表示词序列的长度，$f_j(y_{i-1}, y_i, \boldsymbol{X}, i)$ 是特征函数，λ_j 是第 j 个特征函数的权重系数。

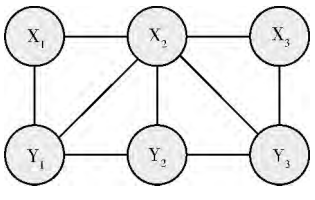

图 1 一个普遍意义的条件随机场

时间信息识别问题可以转化为序列标注问题，其要求是在给定观察序列 $\boldsymbol{X}$ 的条件下，估计产生标注序列 y 的概率。而 CRF 模型可以轻易地将观察序列中的任意特征加入到模型中，从而较好的解决这一问题。

3.2 基于 CRF 的时间信息抽取

基于 CRF 的时间抽取模型如图 2 所示。

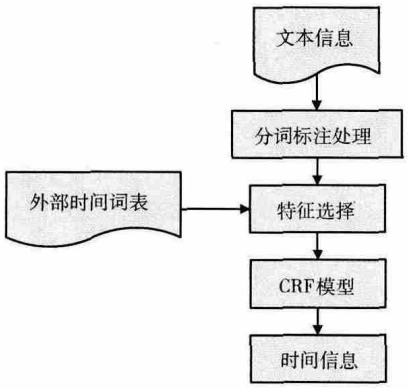

图 2 基于 CRF 的时间抽取

3.2.1 分词和标注处理

在时间信息的识别中，为了充分利用词性和语义特征，分词是必不可少的环节。由于具有比较好的效果，作者使用中科院的 ictclas 分词工具对文本信息进行切词和词性标注[12]。文章使用 B、I、O 标注方法来标记文本中的时间实体，句子中每个词的类型都是 B、I、O 标注中的一种。如果一个时间序列由几个词组成，则 B 表示第一个时间词，第二个以后的都用 I 表示，不是时间词的都用 O 表示。例如："昨天下午 4 点 28 分，十堰到武昌的火车到站"，标注为："昨天/B 下午/I4 点/I28 分/I，/O 十堰/O 到/O 武昌/O 的/O 火车/O 到站/O"。

3.2.2 特征选择

CRF 的训练中，最重要的是特征的选择，在此采用廖先桃等提出的特征模板，特征主要涉及词级特征，包括词、词性、词与其词性的组合和词的上下文特征等。

对时间信息的识别要依赖于时间触发词、时间词的前缀后缀和上下文关联词。根据语料的分词结果，通过程序和人工结合建立时间触发词表（如：立即，当前，马上，等）、前缀词表（如：直到，在，从，等）和后缀词表（如：左右，期间，之前，等）。除了这些词表外，我们还考虑词性、时间词的前词性特征、后词性特征，以及短语的位置特征，即该词是否在句首、是否在句尾等，这些特征见表 1。我们将这些特征都抽取出来后，制作特征模板并用适当的工具训练生成模板文件。

表 1 选取的特征

类别	特征
词性	当前词的词性
	当前词前词的词性
	当前词后词的词性
词表	当前词是否在时间词表中
	当前词前词是否在时间词表中
	当前词后词是否在时间词表中
位置	当前词是否在句首
	当前词是否在句尾

3.3 基于半监督的训练

统计机器学习方法的优点是智能化较高，人工干预较少，而相应的也面临着一些困难，主要就是训练数据不足，可用于命名实体研究的语料也比较缺乏，而且对于不同的领域，往往需要建立不同的语料库，比较耗时耗力，而且在时间上也不够效率。因此，本文采用了自训练（self-training），一种半监督的学习方法来有效利用大量的未作标注的未分类数据，从而提高时间识别在真实文本上的泛化能力。自训练的模式如图 3 所示。

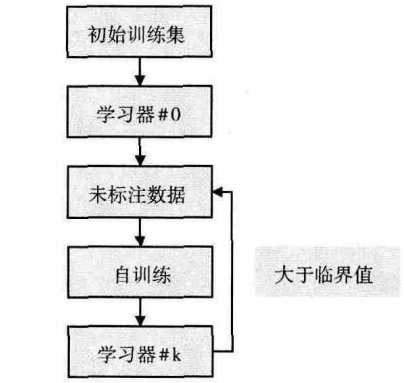

图 3 自训练模式

采用的算法步骤如下：

输入：

初始训练数据集合 D_s，未标注数据集合 D_t，利用 D_s 训练出一个初始学习器 $\#0$，令 $i=0$；

循环部分：用学习器 $\#i$ 对数据集合 D_t 进行预测，在预测结果中取置信度较高（大于某一临界值）的数据为集合 D_t'，令 $D_s=D_s+D_t'$，$D_t=D_t-D_t'$，$i++$ 直到 D_t' 为空这一条件满足；

输出：

n 个学习器 $\#0$、$\#1$、…、$\#k$…

每轮迭代可以得到学习器 k 和新的标注数据集 D_s，临界值需要在实验过程中根据预测结果的多少动态地计算求得，比如初始设定的临界值为 t_0，在实验过程中发现取得的数据集合 D_t' 过少，这样循环的次数就会无限大，在此情况下 t_0 的值就需要设置为一个较小的数，反之则变大。这样经过 n 轮迭代后可以得到 n 个学习器，对这 n 个学习器进行组合形成最终的模型，在本次实验中我们采用选取所有的学习器给予相同的权重的方式。

4 实验结果与分析

4.1 工具及语料

本实验采用的语料来自 2013 年各个门户网站关于舆情部分的新闻，是用武汉大学自然语言处理实验室开发的专门处理舆情新闻的系统来抽取的。共选取了 3000 篇语料，其中 800 篇进行了手工标注，标注结果见表 2。在这 800 篇

中，随机选择 400 篇作为最终的测试语料，其余 400 篇和未标注的 2200 篇总共 2600 篇作为训练语料。CRF 使用 CRF＋＋0.58，使用 Perl 脚本 conlleval.pl 作为评测工具。

表 2　语料标注结果

语料篇数	时间短语数	显性时间数	隐性时间数
800	4031	3218	813

4.2　实验结果

实验结果采用计算精确率（P）、召回率（R）和 F-measure（$F1$）值作为评测标准

$$P=\frac{\text{系统正确识别的时间短语个数}}{\text{系统识别的时间短语个数}}\times 100\%$$

$$R=\frac{\text{系统正确识别的时间短语个数}}{\text{文本中的时间短语个数}}\times 100\%$$

$$F1=\frac{2PR}{P+R}\times 100\%$$

实验分 3 组，分别对显性时间、隐性时间和总体时间进行实验。

4.2.1　规则抽取的实验结果

基于迭代的规则和词典相结合的方法，利用训练语料进行训练，使用其结果对规则和词典进行补充，最终对测试语料的实验结果见表 3。

表 3　基于规则的实验结果

类别	召回率(P)	准确率(R)	$F1$
显性时间	92.61%	96.83%	94.67%
隐性时间	79.45%	91.36%	84.99%
总体时间	89.57%	95.49%	92.44%

从结果我们可以看出，规则的方法对于显性时间信息的抽取效果还不错，可以达到 94.67%的 $F1$ 值，但对于隐性时间，$F1$ 值只有 84.99%，特别是召回率，连 80%都没有达到，所以后续实验中，我们将重点关注隐性时间提升的效果。

4.2.2　半监督方法对时间信息抽取的性能提升

在 2600 篇训练语料中，以 400 篇的标注数据作为初始训练集，剩余 2200 篇为未标注数据，动态地计算临界值的办法，经过循环后，迭代退出得到 7 个学习器，分别对测试语料进行实验，得到的效果如图 4～图 6 所示。

从图 4～图 6 可以看出，基于自训练模式得到的学习器，基本上在到达第 4 轮或第 5 轮迭代时，模型的性能达到最高，然后就在一个小范围的幅度内稳定地波动。同时也可以看到，性能的提升基本上只有 1 个百分点左右，是比较有限的，主要是因为 CRF 特征选取的比较好的情况下，在对时间信息的抽取方面，原系统的性能已相当高，可提升的空间本来就有限。总体而言，半监督的训练对系统性能有一定的帮助。

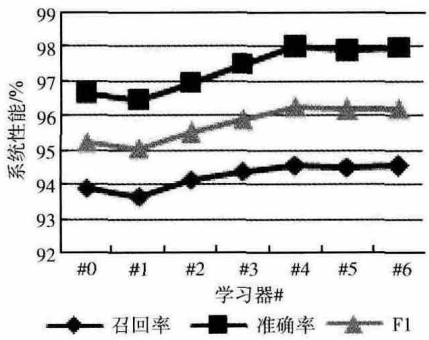

图 4　显性时间抽取的自训练

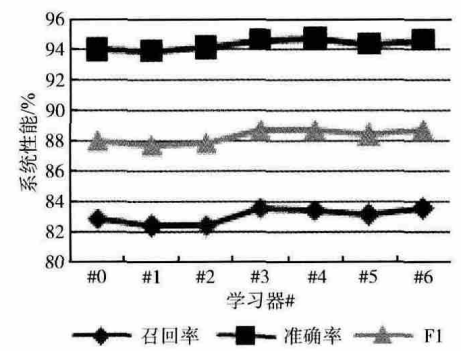

图 5　隐性时间抽取的自训练

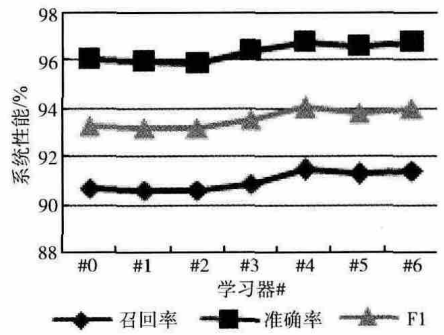

图 6　总体时间抽取的自训练

4.2.3　性能比较

根据上一步得到的实验结果，我们对所有的学习器采用相同权重进行组合，即对于学习器＃0、＃1、＃2、＃3、＃4、＃5、＃6 分别给予 1/7 的权重，再对测试语料进行实验，最终得到的结果见表 4。

表 4　自训练的组合方式的抽取结果

类别	召回率(P)	准确率(R)	$F1$
显性时间	94.55%	98.01%	96.25%
隐性时间	83.39%	94.62%	88.65%
总体时间	91.38%	96.72%	93.97%

对比表 3，我们发现，在显性时间的抽取方面，$F1$ 值略微有些提升；而对于隐性时间，$F1$ 值提升了将近 4 个百分点，效果还是很明显的；总体时间的 $F1$ 值提升的也不

多，是因为语料中隐性时间所占的比重比较小。总体而言，自训练方法还是取得了不错的效果。

5 结束语

通过对中文时间信息的分类和时间信息抽取研究现状的分析，挖掘时间短语在文本中的语言学特征，引出了时间信息抽取的研究思路，确定了用 CRF 的方法，利用 BIO 标注模式将时间识别问题转化为序列标注问题，并通过自学习这样一种半监督的方法对语料进行训练，最终通过对测试语料的实验，取得了较好的效果。对于显性时间的识别，由于 $F1$ 值已经达到了 96.25%，所以提升余地比较小；而对于隐性时间的识别，还是有一定的提升空间的。下一步的工作主要为以下几个方面：①要进一步挖掘隐性时间的各种有效特征，对其进行研究和探讨，以最终提升总体时间的识别率；②改进半监督的学习方法，优化自训练的算法，以得到更好效果的学习器，并最终提升系统性能；③将此种基于 CRF 和半监督训练的方法应用到地名、人名和组织名等其它命名实体抽取的工作中。

参考文献：

[1] Pawel Mazur, Robert Dale. A rule based approach to temporal expression tagging [C] //Proceeding of the International Multiconference on Computer Science and Information Technology, 2007: 293-303.

[2] ZHOU Xiaojia, ZHOU Qingli. The research on the extraction of temporal information from Chinese medical narrative records [C] //Zhejiang Province Ninth Annual Conference Proceedings on Medical Engineering Branch of Medical Association, 2011: 300-305 (in Chinese). [周小甲，周庆利. 中文病历文本中时间信息自动标注 [C] //浙江省医学会医学工程学分会第九届学术年会论文汇编，2011：300-305.]

[3] Chinchor N, Brown E, Ferro L, et al. 1999 Named entity recognition task definition version1.4 [EB/OL]. [2011-08-05]. ftp://jaguar.ncsl.nist.gov/ace/phase1/ne99__taskdef__v1__4.pdf.

[4] Linguistic data consortium, ace (automatic content extraction) Chinese annotation guidelines for TIMEX2 [EB/OL]. [2009-12-08]. http://www.ldc.upenn.edu/Projects/ACE.

[5] Past TAC (text analysis conference) data [EB/OL]. [2011-08-05]. http://www.nist.gov/tac/data.

[6] JIANG Wenzhi, GU Jiaojiao, CONG Linhu. Research on CRF and rules based military named entity recognition [J]. Command Control & Simulation, 2011, 33 (4): 13-15 (in Chinese). [姜文志，顾佼佼，丛林虎. CRF 与规则相结合的军事命名实体识别研究 [J]. 指挥控制与仿真，2011，33 (4)：13-15.]

[7] WANG Feng'e, TAN Hongye, QIAN Yili. Recognition of temporal relation in one sentence based on maximum entrooy [J]. Computer Engineering, 2012, 38 (4): 37-39 (in Chinese). [王风娥，谭红叶，钱揖丽. 基于最大熵的句内时间关系识别 [J]. 计算机工程，2012，38 (4)：37-39.]

[8] Li Fenghuan, Zheng Dequan, Zhao Tiejun. Event recognition based on time series characteristics [C] //Proceedings of Conference on Fuzzy Systems and Knowledge Discovery, 2011: 1807-1811.

[9] SHI Haifeng, YAO Jianmin. Study on CRF-based Chinese named entity recognition [D]. Suzhou: Soochow University, 2010 (in Chinese). [史海峰，姚建民. 基于 CRF 的中文命名实体识别研究 [D]. 苏州：苏州大学，2010.]

[10] LIU Li, HE Zhongshi, XING Xinlai, et al. Chinese time expression recognition based on semantic role [J]. Application Research of Computers, 2011, 28 (7): 2543-2545 (in Chinese). [刘莉，何中市，邢欣来，等. 基于语义角色的中文时间表达式识别 [J]. 计算机应用研究，2011，28 (7)：2543-2545.]

[11] LIAO Xiantao. A study on Chinese named entity recognition [D]. Harbin: Harbin Institute of Technology, 2006 (in Chinese). [廖先桃. 中文命名实体识别方法研究 [D]. 哈尔滨：哈尔滨工业大学，2006.]

[12] WANG Feng'e. Recognition of temporal relation in Chinese texts [D]. Taiyuan: Shanxi University, 2012 (in Chinese). [王风娥. 汉语文本中的时间关系识别技术研究 [D]. 太原：山西大学，2012.]